INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Statistical and Risk Modelling 1

Chapter: Unit 1 & 2

Category: Assignment Solutions

1.

${}_{u}q_x = Pr\left[T_x \le u\right] = Pr[K_x = 0 \text{ and } S_x \le u] = Pr\left[K_x = 0\right] * Pr\left[S_x \le u\right]$ since K_x and S_x are independent. $Pr\left[S_x \le u\right] = \int_0^u 1dx = u$, since uniform distribution.

$Pr\left[S_x \le u\right] = \int_0^u 1dx = u$, since uniform distribution.

Thus, $q_x = u^* q_x$ since $Pr\left[K_x = 0\right] = q_x$

2.

a) Central exposed to risk
Period of exposure is $1.6.2000$ to $25.10.2000$

$$= 30 + 31 + 31 + 30 + 25 = 147\ days \quad = 147/7 = 21\ weeks$$

b) Initial exposed to risk
Period of exposure is $1.6.2000$ to $31.5.2001 = 52$ weeks

3. i) 11 athletes qualified during the period of observation, so the median is the number of events taken to qualify by the sixth athlete to qualify. This is 9 events.

ii) Define *t* as the number of events which have taken place since 1 Jan 2014.
Injured and stopped participating implies recorded *after* the event number reported.

t_j	N_j	D_j	C_j	D_j/N_j	$1-D_j/N_j$
0	23	0	2	0	1
6	21	1	0	1/21	20/21
8	20	2	1	2/20	18/20
9	17	3	0	3/17	14/17
11	14	2	1	2/14	12/14
13	11	3	0	3/11	8/11

The Kaplan-Meier estimate is given by product of **$1-D_j/N_j$**

Then the Kaplan-Meier estimate of the survival function is

t	$S(t)$
$0 \le t < 6$	1
$6 \le t < 8$	0.9524
$8 \le t < 9$	0.8571
$9 \le t < 11$	0.7059
$11 \le t < 13$	0.6050
$13 \le t < 14$	0.4400

iii) The median time to qualify as estimated by the Kaplan-Meier estimate is the first time at which S(t) is below 0.5. Therefore, the estimate is 13 events.

iv) The estimate based on athletes qualifying during the period is a biased estimate because it does not contain information about athletes still participating at the end of the period, or about those who dropped out (injured and stopped participating without qualifying).

The athletes still participating at the end of 2020 have (by definition) a longer period to qualification than those who qualified in the period.

Hence the Kaplan-Meier estimate is higher than the median using only athletes who qualified during the period.

4.

i. complete expectation of life, $\mathring{e}_x$

$$\mathring{e}_x = E[T_x] = \int_0^{\omega-x} {}_tp_x \, dt$$

This represents the integral of the probability of survival at each future age, i.e. the expected future lifetime of a life currently aged x. In other words, this is the expectation of life at age x or how many years a life is expected to live given that it is currently x years old.

ii. The curtate expectation of life

$$e_x = \sum_{k=1}^{\infty} {}_kp_0 = \sum_{k=1}^{\infty} e^{-0.0325k} = \frac{e^{-0.0325}}{1-e^{-0.0325}} = 30.27$$

iii. The probability that a life aged exactly 36 will survive to age 45.

$${}_9p_{36} = \exp\left[-\int_0^9 0.0325\, dt\right] = e^{-0.2925} = 0.7464 \approx 75\%$$

iv. The exact age x representing the median of the life-time T of a new born baby.

The median of the life-time T implies that the probability, ${}_xp_0 = 0.5$

Thus, ${}_xp_0 = 0.5 \Rightarrow \exp(x.-0.0325) = 0.5 \Longrightarrow x = -\frac{\log 0.5}{0.0325} = 21.33$

5. (i) Gompertz Law is a suitable model for human mortality for middle to older ages say 35 and over.

There is evidence that the Gompertz Law breaks down at very advanced ages and therefore 35 to 90 years is acceptable.

(ii) Since $${}_tP_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

Putting $\mu_x = Bc^x$

$${}_tP_x = \exp\left(-\int_0^t Bc^{x+s}\,ds\right)$$

We can write c^{x+s} as $\longrightarrow c^x e^{s\log c}$

$$\int_0^t Bc^{x+s}\,ds = \int_0^t Bc^x e^{s\log c}\,ds = \frac{Bc^x}{\log c}\left[e^{s\log c}\right]_0^t$$

$$\frac{Bc^x}{\log c}\left[e^{s\log c}\right]_0^t = \frac{Bc^x}{\log c}\left[c^s\right]_0^t = \frac{Bc^x}{\log c}\left[c^t - 1\right]$$

If we introduce the auxiliary parameter g defined by log g = -B/ log c, the value of the integral is –log g $c^x(c^t - 1)$ and hence

$${}_tP_x = \exp\left(\log g c^x\left[c^t - 1\right]\right) = \left(e^{\log g}\right)^{c^x(c^t-1)} = g^{c^x(c^t-1)}$$

6.

(i) Female smoker aged 30 at entry.

(ii) $$\frac{h_j(t)}{h_i(t)} = \frac{\exp(-.05)}{\exp(0.1)} = 0.86070$$

Where j is male smoker aged 30 at entry and i is female smoker aged 40 at entry

But s(t) = exp (- $\int_0^t h(s)ds$) hence

$$s_j(t) = (s_i(t))^{0.86070}$$

which implies that

$s_j(t) > s_i(t)$ for all $t > 0$

(iii) $$\frac{h_j(t)}{h_i(t)} = \frac{\exp(0.2)}{\exp(0.05)} = 1.161$$

Where j is male smoker aged 30 at entry and i is male smoker aged 40 at entry

But s(t) = exp (- $\int_0^t h(s)ds$) hence

$$s_j(t) = (s_i(t))^{1.161}$$

Which implies that

$s_j(t) < s_i(t)$ for all $t > 0$

7. i) The contribution of each life to the central exposed to risk is the number of months between STARTDATE and ENDDATE, where

STARTDATE is the latest of (date of 50th birthday, 1 January 2019) and

ENDDATE is the earliest of (date of 51st birthday, date of death, 31 December 2019)

Life	*Date of 50th birthday*	*Date of death*	*Start Date*	*End Date*	*Duration in months between Start and End of observation*
1	01 February 2018	–	01 January 2019	31 January 2019	1
2	01 April 2018	01 October 2019	01 January 2019	31 March 2019	3
3	01 June 2018	–	01 January 2019	31 May 2019	5
4	01 September 2018	–	01 January 2019	31 August 2019	8
5	01 November 2018	15 March 2019	01 January 2019	15 March 2019	2.5
6	01 January 2019	–	01 January 2019	31 December 2019	12
7	01 May 2019	15 December 2019	01 May 2019	15 December 2019	7.5
8	01 July 2019	01 October 2019	01 July 2019	01 October 2019	3
9	01 August 2019	–	01 August 2019	31 December 2019	5
10	01 December 2019	–	01 December 2019	31 December 2019	1
				Total	**48**

Central exposed to risk is the sum of contribution of each of the 10 lives (in number of months) to the observation i.e. 48 months or 4 years.

ii) The total number of deaths during the period of observation is 3. So, the maximum likelihood estimate of the hazard of death is 3/4 = 0.75.

iii)

ALTERNATIVE 1

If the hazard of death at age 50 years is μ_{50} , then

$q_{50} = 1 - p_{50} = 1 - \exp(-\mu_{50})$

$= 1 - \exp(-0.75) = 1 - 0.4724 = 0.5276.$

ALTERNATIVE 2

If the central exposed to risk is E^c_{50}, then if we work in years

$$q_{50} \approx \frac{d_{50}}{E^c_{50} + 0.5 * d_{50}}$$

$$= \frac{3}{4 + 0.5 * 3}$$

= 3/5.5 = 0.5454

8.

i) Types of censoring presents:

• Type I censoring present because the study ends at a predetermined duration of 45 days.

• Type II censoring is not present because the study did not end after a predetermined number of patients had died.

• Random censoring is present because the duration at which a patient left hospital before the study ended can be considered as a random variable.

• Right Censoring is present for those lives that exit before the end of investigation period

ii) The censoring is likely to be informative.

The patients who died were probably recovering less well that patient who discharged from the hospital.

If they had not died, they would likely to remain in the hospital for longer than those who were not censored.

iii) The Kaplan-Meier estimate of the survival function is estimated as follows:

T	n	d	c	d/n	(1 - d/n)	s(t)
0	13					
5	13	1	0	0.0769	0.9231	0.92
7	12	1	0	0.0833	0.9167	0.85
14	11	1	2	0.0909	0.9091	0.77
28	8	1	2	0.1250	0.8750	0.67
35	5	1		0.2000	0.8000	0.54

So the value survival function at end of investigation period is 0.54

Assumptions:

- The censoring happens just after the death
- Ignoring the discharge on any other ground except recovery from illness
- Ignore any admission period before the start of investigation

iv) Comments:

• The survival of a patient from the infection who given treatment is around 50% in light of the answer in c) above.

• However, the hospital excluded the number of deaths who died within two weeks of observation period.

• It also ignores the admission pre investigation period

• It is assuming that the censored patient at the end of investigation will survive for sure.

• Also ignoring the patients being discharged on any other ground like shifting to another hospital etc.

• It claims that 8 out of 10 patients who responded the treatment beyond two weeks would survive.

• So, the claims have to be viewed with respect to above considerations. [3]

9.

a) Under the uniform distribution of deaths assumption:

$\int_0^1 tPx\, dt = \int_0^1 (1 - tqx)\, dt = [t - 0.5t^2 q_x]_0^1$

$= 1 - 0.5q_x$

since $q_x = 0.3$, we have

$m_x = \frac{0.3}{1-.15} = 0.352941$

b) Under the constant force of mortality:

$q_x = 1 - e^{-\mu}$

$\int_0^1 tPx\, dt = \int_0^1 e^{-\mu t}\, dt = \frac{1}{\mu}(1 - e^{-\mu}) = q_x/\mu$

So, $m_x = \mu = -\ln(1 - q_x) = -\ln 0.7 = 0.356675$

10.

i) Under the Cox model each individual's hazard is proportional to the baseline hazard, with the constant of proportionality depending on certain measurable quantities called co-variates. Hence the model is also called a proportional hazards model.

ii)

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + \ldots + b_px_p)$$

iii) The baseline hazard refers to annual policy taken through the Online channel and where premiums are paid by direct debit [1]

iv) The results imply that

$\exp[(\beta_D *1)]/ \exp[(\beta_D *1) + \beta_F*1 + \beta_M*1] = 0.75$ i.e

$\exp(\beta_F + \beta_M) = 4/3$ Eqn 1

$\exp(\beta_D*1) / \exp[(\beta_F *1)] = 1$ Eqn 2

$\exp(\beta_M*1) / \exp[(\beta_D *2)] = 0.75$ Eqn 3

Substituting from (2) into (1) gives

$\exp(\beta_D + \beta_M) = 4/3$

$\exp(\beta_D) * \exp(\beta_M) = 4/3$ Eqn 4

From Eqn 3

$(\exp(\beta_D))^2 *0.75 = \exp(\beta_M)$

So

Substituting in Eqn 4

$\exp(\beta_D) * (\exp(\beta_D))^2 *0.75 = 4/3$

$(\exp(\beta_D))^3 = 1.7778$

$\exp(\beta_D) = 1.2114$

$\beta_D = 0.19179$

$\beta_F = 0.19179$

$\beta_M = 0.0959$

11.

Consider the durations tj at which events take place.
Let the number of deaths at duration tj be dj and the number of insects still at risk of death at duration tj be nj.
At tj = 1, S(t) falls from 1.0000 to 0.9167.
Since the Kaplan-Meier estimate of S(t) is

$$S(t) = \prod_{tj \le t} (1 - \lambda(tj))$$

we must have 0.9167 = 1- $\lambda(1)$,
so that $\lambda(1)$ 0.0833.

Since $\lambda(1)=\frac{d1}{n1}$, then we have $\frac{d1}{n1}$=0.0833
and, since all 12 insects are at risk of dying at tj = 1, we must therefore have d1 = 1 and n1 = 12.
Similarly, at tj = 3, we must have 0.7130 = 0.9167(1- $\lambda(3)$)

so that $\lambda(3)$) = $\left(\frac{0.9167-0.7130}{0.9167}\right)$ =0.222 = $\frac{d3}{n3}$

Since we can have at most 11 insects in the risk set at tj = 3, we must have d3 = 2 and n3 = 9.
Similarly, at tj = 6, we must have 0.4278= 0.7130(1- $\lambda(6)$)

so that $\lambda(6)$) = $\left(\frac{0.7130-0.4278}{0.7130}\right)$ =0.400=$\frac{d6}{n6}$

Since we can have at most 7 insects in the risk set at tj = 6, we must have d6 = 2 and n6 = 5.
Therefore 2 insects died at duration 3 weeks and 2 insects died at duration 6 weeks.

Alternate Solution

t	S(t)	λ(t)	nt	dt	ct
0	1.0000	0	12	0	
1	0.9167	0.0833	12	1	2
3	0.7130	0.22	9	2	2
6	0.4278	0.4	5	2	3

iii) Summing up the number of deaths we have total deaths = d1+d3+d6= 1+2+2= 5.
Since we started with 12 insects, the remaining 7 insects' histories were right censored.

12.

i) **Gompertz Law:**
Gompertz Law is an exponential function, and it is often a reasonable assumption for middle and older ages. It can be expressed as follows:

$\lambda_x = Bc^x$; where, λ_x *is* a force of mortality at age x.

[1]

ii) Substituting, $B = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$; into the Gompertz model,

$\lambda_x = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2).\, c^x$; defining x as duration since 50th birthday.

The hazard can therefore be factorized into two parts:

$\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$, which depends only on the values of the covariates, and
c^x, which depends only on duration.

So, the ration of between the hazards for any two persons with different characteristics does not depend on duration, and so the model is a proportional hazards model. [3]

iii) The baseline hazard in this model relates to a non-smoker female. [1]

iv) For a female cigarette smoker, we have

$X_1 = 0 \; and \; X_2 = 1$ and x = 4

Therefore the hazard at age 54 is given by

$\lambda_x = \; exp(\beta_0 + \beta_1 . 0 + \beta_2 . 1) . c^4$

= exp(-4+0.65)x1.05^4

= 0.0351x1.2155

= 0.04266

v) The hazard for a non-smoker at duration, ‘*s*’ is given by the formula

$$\lambda_s = \exp(\beta_0 + \beta_1 X_1).\, c^s$$

The hazard for a smoker at duration, ‘t’ is given by the formula

$$\lambda_t^* = \exp(\beta_0 + \beta_1 X_1 + 0.65).\, c^t$$

If the smoker’s and non-smoker’s hazards are the same, then

$$\lambda_s = \lambda_t^*$$

i.e., $\exp(\beta_0 + \beta_1 X_1).\, c^s = \exp(\beta_0 + \beta_1 X_1 + 0.65).\, c^t$

i.e. $c^s = \exp(0.65).\, c^t$

i.e. $c^{s-t} = \exp(0.65) = 1.9155$

Since, c = 1.05

Hence, $1.05^{s-t} = 1.9155$

So, s-t = ln(1.9155)/ln(1.05) = 0.65/0.04879

s-t = 13.32

Hence, when the two hazards are equal, the non-smoker is approximately years older than the smoker.

13.

i) (Let $P'_x(t)$ be the number of policies inforce aged x nearest birthday at time t .
Also, let $P_x(t)$ be the number of policies inforce aged x last birthday at time t
Let E_x^C refers to the central exposed to risk at age label x respectively.

$E_x^C = \int_{t=0}^{2} P'x(t)dt$

Assuming that $P'_{56}(t)$ is linear over the year (2015,2016) and (2016,2017), we can approximate the exposure as follows

$E_{56}^c = ½*(P'_{56}(2015)+ P'_{56}(2016))+ ½*(P'_{56}(2016)+ P'_{56}(2017))$

$= ½*P'_{56}(2015)+ P'_{56}(2016)+½*P'_{56}(2017)$

Since, the number of policyholders aged label 56 nearest birthday will be between 55.5 and 56.5 i.e. between age label 55 last birthday and 56 last birthday. Assuming that the birthdays are uniformly distributed over the calendar year:

$P'_{56}(2015)$	=	$½*(P_{55}(2015)+ P_{56}(2015))$
	=	20050
Similarly,		
$P'_{56}(2016)$	=	$½*(P_{55}(2016)+ P_{56}(2016))$
	=	20800

And,

$P'_{56}(2017)$	=	$½*(P_{55}(2017)+ P_{56}(2017))$
	=	19250
E_{56}^c	=	½*20050+20800+1/2*19250
	=	40450
μ_{56}	=	d_{56}/ E_{56}^c
	=	1380/40450
	=	0.0341

Deriving the force of mortality for age 57 as above:

$$P'_{57}(2015) = ½*(P_{56}(2015)+ P_{57}(2015))$$
$$= 19850$$

Similarly,

$$P'_{57}(2016) = ½*(P_{56}(2016)+ P_{57}(2016))$$
$$= 20900$$

And,

$$P'_{57}(2017) = ½*(P_{56}(2017)+ P_{57}(2017))$$
$$= 17500$$

$$E_{57}^{c} = ½*19850+20900+1/2*17500$$
$$= 39575$$

$$\mu_{57} = d_{57}/ E_{57}^{c}$$
$$= 1420/39575$$
$$= 0.03588$$

dx is deaths aged x nearest birthday on the date of death. So the age label at death changes with reference to life year. Therefore the age at the middle of life year is x and estimates μ_x.

[6]

ii) We can estimate the initial rates of mortality using the estimated values of μ from part (i) and the following formula

$$q_{55.5} = 1- \exp(-\mu_{56})$$
$$= 0.0335$$

And

$$q_{56.5} = 1- \exp(-\mu_{57})$$
$$= 0.0352$$

[2]

14.

i) The hazard function for getting married is given by:
$\lambda(t,Z) = \lambda_0(t)\exp[0.3 Z_1 + 0.2 Z_2 + 0.3 Z_3 + 0.5 Z_4 - 0.1 Z_5 + 0.7 Z_6 + 0.5 Z_7 - 0.4 Z_8]$

Where
$\lambda_0(t)$ = baseline hazard at time t since looking for the life partner.

$Z = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8)$
Z_1 = 1 if female, 0 if not.
Z_2 = 1 if location = Non Metro, 0 if not
Z_3 = 1 if profession = Service, 0 if not
Z_4 = 1 if profession = Business, 0 if not
Z_5 = 1 if profession = Social Service, 0 if not
Z_6 = 1 if Age Band = 20-25, 0 if not
Z_7 = 1 if Age Band = 25-30, 0 if not.
Z_8 = 1 if Age Band = 35-40, 0 if not.

ii) People most likely to stay single with the lowest hazard function.

The probability that a person who has been looking for a life partner for one year will stay single for next 2 years is:

exp (-integral$|^3_1\lambda(t,Z)\, dt$) [1

If the person is a female, profession as a social service and aged 37, the probability is:
$P_F = \exp[-e^{0.3Z_1 - 0.1Z_5 - 0.4Z_8}$ integral$|_1^3\lambda_0(t)dt$]
$P_F = \exp[e^{-0.2}$ integral$|_1^3\lambda_0(t)dt]$,
Let A = exp [integral$|_1^3\lambda_0(t)dt$]

$P_F = Ae^{-0.2} = 0.3$
A = 0.2298

If the person is a male, working as a businessman and aged 24, the probability is:
$P_M = \exp[-e^{0.2Z_2 + 0.5Z_4 + 0.7Z_6}$ integral$|_1^3\lambda_0(t)dt$]
Let A = exp [integral$|_1^3\lambda_0(t)dt$]
$P_M = A\, e^{1.4} = 0.00257$

15.

i) **Advantages of central exposed to risk.**

Two advantages of central exposed to risk over initial exposed to risk are:

1. The central exposed to risk is simpler to calculate from the data typically available compared to the initial exposed to risk. Moreover, central exposed to risk has an intuitive appeal as the total observed waiting time and is easier to understand than the initial exposed to risk.

2. It is difficult to interpret initial exposed to risk in terms of the underlying process being modelled if the number of decrements under study increase or the situations become more elaborate. On the contrary the central exposed to risk is more versatile and it is easy to extend the concept of central exposed to risk to cover more elaborate situations.

ii) **Calculation of exposed to risk.**

Rita

Rita turned 30 on 1 October 2009, when she was already married. She died on 1 January 2010, 3 months after her 30^{th} birthday.

Thus. Rita's contribution to central exposed to risk = 3 months

And contribution to initial exposed to risk = 1 year

Sita

Sita turned 30 on 1 September 2011, when she was already married. Time spent under investigation, aged 30 last birthday by Sita was I September 2011 - 31 August 2012.

Thus. Sita's contribution to both central and initial exposed to risk is 1 year.

Nita

Nita turned 30 on 1 December 2009 and married 2 months later. Therefore, she joined the investigation of married women on 1 February 2010. She divorced 9 months later, when she would be censored from the investigation of married women.

Thus, Nita's contribution to both central and initial exposed to risk is 9 months.

Gita

Gita got married on 1 June 2011, at which time she was already past her 31st birthday. Therefore, she has spent no time during the investigation period as a married woman at age 30 last birthday.

Thus, her contribution to both central and initial exposed to risk is nil.

iii) **Total exposed to risk**.

Hence, total exposed to risk is:

Central exposed to risk = 0.25 + 1 + 0.75 + 0 = 2 years.

Initial exposed to risk = 1 + 1 + 0.75 + 0 = 2.75 years

From the results above, it can be seen that the central exposed to risk is 2 years and the initial exposed to risk is 2.75 years. The approximation would suggest that the initial exposed to risk should be 2.5 years.

However, this is not a good approximation for the data provided as the approximation is based on the assumption that deaths would be evenly spread and thus can be assumed to occur half way through the year, on average. This also relies on an implicit assumption of a reasonably large data set. In the data above, there were only 4 lives, which is not statistically significant. Moreover, there was only one death, which occurred 3 months after the 30thbirthday. As a result of the statistical sparseness in the data, the approximation is seen not to work very well.